1. **Introduction**

The silicon photomultiplier (SiPM), also known as a multi-pixel photon counter (MPPC), is a light sensor consisting of a matrix of thousands of single-photon avalanche diodes (SPADs) called microcells, or pixels [1]. All the pixels are connected in parallel to the same readout through quenching resistors. When a photon is absorbed by a pixel, a significant electronic pulse is generated, and the output can be as large as tens of micro-amps from a single pixel. This occurs a result of the avalanche that occurs at the SPAD junction. The electrical charge generated from absorbing the photon depends only on the junction voltage and capacitance, and does not depend on the incident photon’s energy. This state is called the Geiger mode. Thus, a semi-analog output is obtained, and its size depends on the number of photons that interacted with the SiPM sensor. The SiPM is unique because, due to the high intrinsic (internal( gain, it is the first solid-state device able to identify a single photon at room temperature [2]. The SiPM was developed in the early 2000s in a collaboration between the Moscow Engineering Physics Institute (MEPhI) and Pulsar. Since approximately 2005, many companies recognized the new technology’s potential and began to develop and manufacture poly-silicon-based SiPM components. Among the leading companies in the field are Hamamatsu, SensL, FBK, STMicroelectronics, KETEK, and others. The SiPM replaced older light sensors in many applications with low luminosity (signal/information). Among other applications, SiPM components serve as light sensors in range detection, light detection and ranging (LIDAR) systems, which are used in the navigation systems in autonomous cars, quantum informatics [3], for spectroscopy in physical experiments, and more. In addition, SiPM components function as light sensors in scintillation-based radiation detectors. They are used in medical imaging devices such as positron emission tomography (PET) and single photon emission tomography (SPECT), nuclear safety products (homeland security), radiation detectors, and detectors for space exploration. Its advantages over other light sensors, such as the photomultiplier tube (PMT), are immunity to electromagnetic interference (an especially important advantage in the context of medical imaging devices), low operating voltage (tens of volts), compactness, and cost. Along with the above, the SiPM is an array of solid-state devices that are reverse-biased by high voltage and therefore it experiences significant dark currents, some of which result from tunneling. The tunneling phenomenon becomes more dominant for low and cryogenic temperatures. The functioning of the SiPM at such temperatures is especially relevant for SiPM-based detectors on satellites and in particle research. In both cases, detectors are exposed to extremely low temperatures, where the tunneling process is dominant. In this work, we will review the SiPM and its features, as well as tunneling and its effect on the dark current of the SiPM component. We will also present the use of tunneling in the SiPM configuration, to expand the dynamic range of the device.

1. **Introduction to SiPMs**

The SiPM consists of a matrix of thousands to tens of thousands of pixels, or single-photon avalanche diodes (SPADs), each one connected through an RQ resistor to common-mode bias voltage. The analog SiPM output is a junction connecting all the SPAD cell anodes. Figure 2 shows a schematic diagram of the SiPM structure. The load resistor, RL, is not part of the device, rather it is externally connected.

The SPAD cells are also called Geiger-mode avalanche photodiodes (GM-APD). They are solid-state devices that utilize the avalanche in the depletion layer to create an internal gain mechanism. As early as the 1960s, avalanche photodiodes (APD) were studied at voltages close to breakdown voltages. In the 1990s, for the first time, APDs were introduced that operated in Geiger mode above the breakdown voltage; in this state, any discharge produced was “digital” with uniform charge, independent of the incident photon’s energy. These components are called GM-APD or SPADs. Figure 3 shows the division between the working ranges of the SiPM (as a component based on SPADs) and the APD.

Besides the physical operating voltages, SPADs differ from APDs in their internal structure, internal gain (hundreds for APDs, in comparison with 106 for SPADs), and signal-to-noise ratio. The avalanche process also fades on its own in APDs, without the need for quenching.

Today, SPADs are made from silicon, mostly using CMOS technology. Each SPAD pixel is in fact a p-n junction designed to be biased above the breakdown voltage of the junction. The field generated in the depletion area is on the order of 105 V/cm. An electron generated in the depletion area creates a self-sustaining avalanche of electrons under the high electrical field. The current generated causes a voltage drop at the RQ resistor, which leads to the voltage on the cell falling to the breakdown voltage and the subsequent fading of the avalanche process. Figure 4 portrays the avalanche process in a single pixel.

In the first stage, the SPAD is biased at voltage VB, above the breakdown voltage VBR. At this stage, no current flows through the pixel. During the formation of a free charge carrier in the depletion layer, the charge carrier drifts in the high electrical field, either as a result of an incident photon and electron release as a part of the photoelectric effect, or as a result of thermal excitation of an electron. The high speed of the charge carrier causes the release of additional charge carriers that also drift, and in turn release even more charge carriers, thereby generating an intrinsic charge gain. The current flows through an RQ resistor (typically several hundred kiloohm), causing a junction voltage drop close to the breakdown voltage and halting the avalanche process. The cell voltage increases to the bias voltage again, a reset process, with a time constant determined by multiplying the combined pixel capacitance and the parasitic capacitance of the RQ resistor at the output load of the SiPM, as described in Equation (1).

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where

CSPAD is the capacitance of the p-n junction.

CQ is the parasitic capacitance of the RQ resistor.

Figure 5A shows the structure of a single SPAD unit. Figure 5B displays an X-ray image of a SPAD unit. The RQ resistor can be identified around the light-sensitive region.

An integrated RQ resistor, shown in Figure 5B, allows us to obtain the Fill Factor (FF), an optimal ratio between the light-sensitive area of the sensor and the total area.

1. **Breakdown voltage**

The breakdown voltage, VBD, is the voltage at which the multiplying of charge carriers sharply increases, as a result of the interaction with the initial charge carrier. VBD is the boundary between the linear multiplying range that characterizes APDs and the avalanche range typical of SPADs. For SiPMs in which an RQ resistor is implemented outside the device, a very sharp increase in current and breakdown voltage can be identified relatively easily. In SiPMs in which the RQ resistor is integrated, the internal gain is smaller by an order of about three (from 108 to about 105); in such a situation, it is necessary to perform a more complex analysis to accurately determine the breakdown voltage. The precision of this voltage is necessary for obtaining uniform gain when working with an array of SiPMs. Breakdown voltage, VBR, depends mainly on the doping level of the junction and the temperature. The breakdown voltage can also be defined by the “ionization integral”, where the breakdown voltage is the voltage at which the integral across the depletion layer is equal to 1, as displayed in Equation 2:

|  |  |  |
| --- | --- | --- |
|   |  | (2) |

Where

αp , αn, are the ionization constants of electrons and holes, and

W is the width of the depletion layer.

Generally, the ionization constant values vary between different models [4], and demonstrate an exponential relationship to the electric field at the junction. Figure 6 shows ionization constants of various materials as a function of the electrical field. It can be seen that for silicon, the ionization constant for electrons, αn (solid line), exceeds the ionization constant for holes, αh (dotted line), in every range.

The breakdown voltage VBR rises with the temperature. This increase can be explained as follows: when a charge carrier drifts in the high electric field, collisions with the atoms in the lattice occur. In the first approximation, the threshold energy at which the charge carrier creates another electron-hole pair is ETH = 3/2Eg , assuming identical mass for the electron and hole. Every time the initial, accelerated charge carrier collides with a phonon or atom with less energy than the threshold energy, ETH, it loses some of its energy without creating secondary charge carriers. The rise in temperature increases the probability of collisions, and therefore the average energy for collisions decreases. This suggests that the electric field (and therefore the voltage) must be increased at the junction to obtain a self-sustaining avalanche, as required in Geiger mode. Also, the width of the depletion layer affects the breakdown voltage. It would seem that a large depletion layer leads to more collisions at the junction, and therefore enables a lower required critical field for switching to Geiger mode. However, since the breakdown voltage is the integral of the field over the entire junction width, widening the depletion layer actually increases the breakdown voltage. When the depletion layer expands, so does the effect of the temperature on the breakdown voltage. With the occurrence of an avalanche, a peak current Lf is formed as described in Equation 3.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Voltage VOV is the device voltage beyond the breakdown voltage. Equation 4 presents the voltage on the junction VF, at the peak of the avalanche process:

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Where RD is the internal resistance of the pixel. When the voltage on the junction falls to voltage Vf, the avalanche ceases to perpetuate itself.

A rise in the electric field near the diode terminals causes a higher multiplication factor and a reduced breakdown voltage near the edges [5]. To minimize this phenomenon, which affects the gain uniformity, a guard ring is used. The structure is made using p-well technology, designed to reduce the electrical field at the edges, and thus prevent the beginning of a premature avalanche in the area. An example structure and simulation of the electrical field with a guard ring is shown in Figure 7. The disadvantage here is the reduction of the active area by 1-2 μm and consequently lower FF. This is rare in SiPMs based on large pixels (for example, 50x50 μm2), but is significant in components based on smaller pixels (10x10 μm2).

1. **Internal Gain**

Internal gain is defined as the area under the output pulse of the SiPM, i.e. the total charge at the output for a single avalanche. The gain is given as

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Where qe is the electron charge. In general, the gain is linear in the first approximation. Factors affecting the linearity of the SiPM are cell uniformity, both in terms of doping uniformity and in terms of CQ values. Similarly, the dynamic range of the SiPM is affected by the density of the cells per unit area and the recovery time (reset) of the cell. The greater the number of cells per unit area and the smaller the “dead time”, in which the cell is not sensitive, the greater the dynamic range of the SiPM gain relative to light intensity. The gain is on the order of 105-107 and produces a detectablesignal for a single photon, beyond the noise level. This has consequences for the power requirements and noise level of electronics using the SiPM, in comparison with other light sensors. The excess noise factor (ENF), noise added as a result of the gain variance in the various pixels, is defined according to Equation 6.

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

Where

σG - standard deviation of gain between SPAD cells.

<G> - average gain in the various cells.

The internal capacitances CSPAD and CQ depend on the size of the SPAD. Therefore, a smaller cell provides a smaller gain, and therefore reduces the signal to noise ratio (SNR). At the same time, a small cell reduces the noise (as described in Chapter 3), and also reduces the reset time. Reducing the reset time also contributes to the reduction of the SNR by reducing the signal integration time, thereby reducing the noise integration. Figure 8a shows the pixel gain and signal amplitude at the output, 8b, obtained from an FBK SiPM measuring 1x1 mm2. As can be seen for increasing VOV voltages, the linear gain is impaired [6]. This phenomenon is a result of penetration of the depletion layer into the device substrate, the epitaxial layer. Penetration of the depletion layer leads to a decrease in diode capacitance, CSPAD, which reduces gain in accordance with Equation 5.

1. **Signal Shaping**

The signal amplitude is not solely dependent on the gain. The reason for this is that it is possible to separate the output signal into two separate components, a fast component and a slow component [7]. The fast component results from the discharge path through the parasitic capacitance CQ, while the slow path results from the charge current of the reset process through resistor RQ. Figure 9 shows an electrical diagram of an SiPM with N pixels. The left rectangle is the schematic diagram of the SPAD in which the avalanche occurs, the middle rectangle represents the passive cells, and the right rectangle describes the parasitic capacitance of the grid connecting the pixels. The red path is the fast current supply path at the beginning of the avalanche. The blue path is the slow path, the cell reset. Equations 7-9 detail the time constants of the output pulse.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |
|  |  | (8) |
|  |  | 9)) |

Where

RL is the SiPM output load.

RD is the internal resistance of a single cell.

Ctotal is the total output capacitance of the pixels.

The general signal form is given according to Equation 10:

|  |  |  |
| --- | --- | --- |
|  |  | 10)) |

Where Qtotal is the total charge released from all component pixels.

Similarly, the output signal depends on the load and bandwidth of the connected electronics. Figure 10 shows different output signals, depending on the bandwidth of the front-end amplifier (FEA) which receives them. For the FBK device, the bandwidth of the high-speed component is in the range of 50-250 MHz. Therefore, for an FEA with bandwidth of 20 MHz, only the slow component of the signal remains.

**4. Detection Efficiency**

The detection efficiency of the SiPM, or the photon detection efficiency (PDE), is the probability that an incident photon will indeed produce an avalanche in an SiPM cell, or in other words, it is the ratio of the number of incident photons to the number of photons detected. The PDE depends on both the voltage supplied to the SiPM beyond the breakdown voltage, VOV, and the wavelength of the incident photon, λ. The PDE value is described in Equation 11.

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

Where

QE is quantum efficiency,

AIP is avalanche initiation probability, and

FF is the fill factor -- geometric efficiency.

The quantum efficiency, QE, is the probability of an incident photon entering the detector (without being reflected, for example) and being absorbed in the active region. The probability of an avalanche being initiated, AIP, depends on the intensity of the electric field and the location of charge carrier formation in the cell. The FF is the probability that the photon will reach the active area in the array and not land between two pixels or in the area of the guard ring. Since the low field width around the guard ring depends on the voltage, the FF also depends on the VOV voltage. Additionally, since the field near the edges depends on the photon absorption depth in the SPAD, the FF also depends on the wavelength of the incident photon.

The PDE based on a p-on-n junction differs from the PDE based on an n-on-p junction. This is because the absorption location of the photon in the cell depends on the wavelength of the incident photon [8], as seen in Figure 11.

While red-green light (500-780 nm) is absorbed within the high field area (assuming a junction width of hundreds of nanometers), blue light (450-490 nm) is absorbed close to the detector before the junction. As shown in Figure 6, electrons have a higher ionization constant than holes. Since the field directions are opposite for p-on-n and n-on-p structures, the p-on-n structure has higher PDE values for blue light wavelengths, while the n-on-p structure yields a higher detection probability for red-green light. Figure 12A shows PDE curves as a function of wavelength for p-on-n and n-on-p SiPM configurations, for epitaxial layers with width of 4 μm and 8 μm. Figure 12B shows quantum efficiency curves as a function of incident light wavelength for various substrate widths.

From Figure 12a, it can be seen that the p-on-n-based device responds better to the blue light range. This is because for a p-on-n junction, interaction above the junction results in electron drift towards it, with a higher ionization constant. This is in contrast to higher wavelength light, for which the interaction occurs below the junction and is formed by drifting holes.

1. **Noise Factors**

The noise in SiPM components can be divided into two main groups:

1. Primary noise or dark current. The source of this noise is charge avalanches created by thermally generated carriers, or carriers generated as a result of tunneling in the high electric field [9].
2. Secondary noise or excess noise. Secondary avalanches are caused as a result of a primary event. The secondary noise is also classified into two types: afterpulsing, an avalanche caused in the same SPAD in which the primary event occurred, and crosstalk, an avalanche caused in cells near the pixel in which the primary event occurred.

The SiPM output is also proportional to the charge generated from the secondary noise factors mentioned above. The excess charge factor (ECF) is defined by the ratio of the total charge to the secondary charge, as defined in Equation 12:

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

Where

QOut is the average total charge at the SiPM output.

QPrimary is the total charge due to noise, afterpulsing, dark noise, and crosstalk events.

Dark Current

The dark current is composed of avalanches that were not created from photon interaction in the pixel of the initial avalanche or in a neighboring pixel. Rather, the dark current results from charge carriers generated as a result of thermal transfer or tunneling in the substrate layer of the cell, or at the depletion layer edge adjacent to the junction. The dark current is the primary noise source in SiPM components. It increases linearly with VOV [11] (assuming negligible added effect from crosstalk and afterpulsing). The rate of dark current events, the Dark Count Rate (DCR), can be calculated when the measurement is performed in dark conditions for a given operating point (bias voltage and temperature), as

|  |  |  |
| --- | --- | --- |
|   |  | (13) |

Where

ID is the SiPM current at the given operating point.

G is the SiPM gain at the measured operating point.

Figure 13 shows DCR as a function of voltage, for SiPM components made by different manufacturers with pixels of various sizes. It can be seen that in some of the components, the function becomes parabolic in high voltage. This is due to the afterpulsing and crosstalk stemming from the initial dark current event. Additionally, the figure demonstrates that the larger the pixel area, the greater the DCR. This is because for larger pixels, the fill factor or the effective detection area increases, while the dead area, generated by the RQ quenching resistor and the separation between the cells, decreased.

At room temperature, the dark current is created mainly in Shockley-Read-Hall (SRH) recombination. A trap is a situation in which electron-hole pairs pass from the conduction band to the valence band by “local energy states” generated in the forbidden gap by lattice defects. The transition between states is based on thermal energy exchange through a phonon. The noise frequency, DCR, is dependent on the quality of the epitaxial layer lattice. Additionally, the DCR depends on the depletion layer volume, i.e. the size of a single pixel and the width of the depletion layer. Generation of SRH pairs is also affected by the high electric field, and is expressed in processes such as the Pole-Frenkel effect, in which electron-hole pairs are created in a non-conductive material, induced by a high electric field or thermally-enhanced trap-assisted-tunneling, which will be explained at length in the following chapter. All these phenomena can be mitigated by reducing the electric field, i.e., by operating at a lower VOV voltage. At higher temperatures, the diffusion current from neutral areas adjacent to the depletion level can also increase the DCR. At low temperatures, the DCR rate decreases significantly by a factor of two for every 10⁰C. At the same time, at lower temperatures, pairs generated from tunneling becomes dominant, and the temperature dependence becomes negligible. This is especially significant for applications designed for low temperature conditions, such as particle detectors and detectors mounted on satellites or research balloons.

The corner temperature, in which the dominant process for dark current generation passes from thermal production to tunneling, varies between SiPM structures, but mainly depends on the size of the junction’s electrical field at breakdown voltage. For a larger electric field, the tunneling is more dominant. Figure XXX shows DCR curves for mm2 as a function of VOV voltage for various temperatures. The DCR rate can be seen to decrease with temperature. Additionally, beginning with a VOV of 2V, it can be seen that the DCR linearly increases with VOV.

Afterpulsing

Afterpulsing is secondary noise resulting from trapping and delayed release of charge carriers in the high field. The trapped charge carrier release time has an exponential distribution, and therefore the probability of an afterpulsing event declines with greater distance from the main event. When the charge carrier is released, it causes a separate avalanche. The probability of afterpulsing depends on the number of traps, the area of high field, and the ratio of the release time constant from the traps to the avalanche reset time. From this, it can be understood that it is possible to decrease the frequency of afterpulsing by increasing the reset time, but this is at the expense of the dynamic range of the SiPM, as it then saturates at a lower incident photon rate. In addition, a longer pulse requires a longer integration time, which increases the electronic noise collected by the electronics connected to the SiPM.

Another configuration of afterpulse generation is by optical induction. During the avalanche, secondary photons are formed by charge carriers hitting lattice atoms. These photons can be absorbed in the neutral area, outside the depletion layer. Absorbing the photon creates a charge carrier that can diffuse to the depletion layer and initiate a secondary avalanche in the same SPAD where the original avalanche was formed. Figure 15 shows an optically-induced afterpulsing event. For such cases, the lifetime of the charge carrier in the epi-layer corresponds to the lifetime of the charge carrier release from the trap. This time varies from several nanoseconds to hundreds of nanoseconds. In order to reduce optically-induced afterpulsing, a substrate with a lifetime shorter than the reset time of the SPAD should be used. In such a situation, the charge carrier contribution to the avalanche would be negligible. Another possible method to reduce optically-induced afterpulsing is creating another p-n junction that blocks the charge carriers from diffusing to the depletion layer.

Crosstalk

As described above, secondary photons are generated during the avalanche process. The quantity of photons was estimated at about 3×10-5 for every avalanche charge carrier [7]. The photon emission from the avalanche is an isotropic process, i.e. the photons distribute in various directions at equal probability. These photons can cause an avalanche in pixels neighboring the pixel in which the initial interaction occurred. Such a response is called direct optical crosstalk, and it is demonstrated by an increased output signal relative to the optical signal. Another form of optical crosstalk is delayed optical crosstalk. It is caused by the formation of electron-hole pairs in the neutral area outside the depletion layer. Again, as in the case of afterpulsing, the charge carriers form a result of photon absorption in the substrate. The charge carriers diffuse and some reach the depletion layer, where they may create a secondary avalanche with a delay of a few nanoseconds to microseconds after the initial pulse. Figure 15 illustrates the path of the photons and charge carriers.